Reliability of Marine Structures Program

THE FITS ROUTINE: FITTING DISTRIBUTIONS TO MULTIPLE DATABASES AND ESTIMATING COMBINED EXTREMES

Steven R. Winterstein

Civil Engineering Department, Stanford University

Report No. RMS–19

June 1996

Department of CIVIL ENGINEERING
STANFORD UNIVERSITY

THE FITS ROUTINE: FITTING DISTRIBUTIONS TO MULTIPLE DATABASES AND ESTIMATING COMBINED EXTREMES

Steven R. Winterstein

Civil Engineering Department, Stanford University

Report No. RMS–19

June 1996

Abstract

This report documents the usage of the routine FITS, which provides automated fits of various analytical, commonly used probability models from input data.

This routine is intended to complement the previously distributed routine, FITTING, documented in RMS Report 14 (Winterstein et al, 1994). The FITTING routine implements relatively complex, four-moment distribution models, whose parameters are fit with numerical optimization routines. While these four-moment fits can be quite useful and faithful to the observed data, their complexity can make them difficult to automate within standard fitting algorithms.

In contrast, the routine FITS is intended to provide more robust (lower moment) fits of simpler, more conventional distribution forms. For each database of interest, the routine estimates the distribution of annual maximum response, based on the data values and the duration, T, over which they were recorded. To focus on upper tails of interest, the user can also supply an arbitrary lower-bound threshold, x_{low}, above which a shifted distribution model—exponential or Weibull—is fit. (In estimating the annual maximum response, the program automatically adjusts for the decreasing rate of response events as the threshold x_{low} is raised.)

Note also that FITS permits modelling of multiple response databases which may represent different statistical populations. Examples include (1) wave or response statistics due to waves from different directions, or (2) responses of Gulf of Mexico structures due to hurricanes alone, eddy currents alone, or combined hurricane-eddy events. By including the relative occurrence rates and response distributions for each population, FITS estimates the total contribution to the annual maximum response.

Contents

1 Distribution Fitting: General Methodology

This report documents the usage of the routine FITS, which provides automated fits of various analytical, commonly used probability models from input data. This routine is intended to complement the previously distributed routine, FITTING, documented in RMS Report 14 (Winterstein et al, 1994). The FITTING routine implements relatively complex, four-moment distribution models, whose parameters are fit with numerical optimization routines. While these four-moment fits can be quite useful and faithful to the observed data, their complexity can make them difficult to automate within standard fitting algorithms.

In contrast, the routine FITS is intended to provide more robust (lower moment) fits of simpler, more conventional distribution forms. For each database of interest, the routine estimates the distribution of annual maximum response, based on the data values and the duration, T, over which they were recorded. It can also estimate the annual maximum response distribution due to multiple types of response "events," intended to reflect different statistical populations. Examples include waves from different directions, or response events due to hurricanes, eddy currents, or combined hurricane-current events.

Specifically, FITS first estimates the individual response CDFs $F_i(x)$=P[Outcome $< x$ in database i] from the data in each database. It then predicts the corresponding distribution of annual maximum response by assuming that events from each database occur in independent, Poisson fashion:

$$P[\text{ Annual max } > x] = G_{ann}(x) = 1 - \exp[-\sum_i \nu_i G_i(x)]\,; \quad G_i(x) = 1 - F_i(x) \tag{1}$$

Here ν_i is the mean annual rate of events in database i, estimated from the observed number per year (possibly over a user-defined threshold; see below).

1.1 Available Distribution Types

Specific distributions currently included in **FITS** to estimate $F_i(x)$ include the following, as catalogued by the distribution index IDIST:

IDIST=1: Normal Distribution

IDIST=2: Lognormal Distribution

IDIST=3: Exponential Distribution

IDIST=4: Weibull Distribution

IDIST=5: Gumbel Distribution

IDIST=6: Shifted Exponential Distribution

IDIST=7: Shifted Weibull Distribution

The distributions IDIST=1 through 5 are all fit to statistical moments of all available data. The single-parameter exponential preserves only the mean m_x of the data, while the normal, lognormal, Weibull, and Gumbel preserve both the mean and standard deviation σ_x estimated from the data. The shifted versions of the exponential and Weibull distributions (IDIST=6 and 7) employ a user-defined lower threshold x_{low}, and fit standard exponential/Weibull models to $x - x_{low}$ based on observed moments. (In estimating the annual maximum response, the program automatically adjusts for the decreasing rate of response events as the threshold x_{low} is raised.)

The result aims to provide the user with a suite of smooth probability models, to be fit throughout the body of the available data. It does not directly address various special topics of data fitting; e.g., selective tail fitting, fitting bimodal models to hybrid data, etc. Some of these issues can be addressed, in a limited way, through the use here of the *shifted* models (IDIST=6 and 7). In this way the user can focus the distribution modelling resources on the extreme response levels of interest.

More specific tail-fitting procedures have not been given here, because optimal use of these tends may be rather problem-specific. In the same vein our extremal models are limited here to so-called "Type I" behavior, leading to (shifted) exponential distributions of peaks over a given threshold and to Gumbel distributions of annual maxima. Type II and III distributions are ill-suited to our moment-fits, due to potential moment divergence (Type II) or to the difficulty in predicting truncated distributions (Type III) from moment information.

1.2 Limitations

Several parameters have been assigned maximum values in the routine FITS. These include the upper limits

- NFMAX, the maximum number of files (databases) to be fit, has been set to 3.
- NMAX, the maximum number of data per database, has been set to 32768.

Both of these limits have been set in PARAMETER statements in the main driver program to FITS. These are rather arbitrarily selected limits, and can be reset by the user without fundamental consequence.

2 Distribution Fitting: Routines

The fitting algorithm calls the following set of subroutines:

CALMOM: Estimates the mean m_x, standard deviation σ_x, skewness α_3 and kurtosis α_4 from an input set of data. These are based on unbiased estimates of the cumulants k_1=m_x, k_2=σ_x^2, k_3=$\alpha_3\sigma_x^3$, and k_4=$(\alpha_4-3)\sigma_x^4$. If the user includes an optional lower limit x_{min}, moments of the shifted variable $(x - x_{min})^+$=max$(0, x - x_{min})$ are estimated.

DISPAR: Based on the sample moments estimated in CALMOM, DISPAR seeks a consistent set of distribution parameters. The interpretation of these parameters depends on the distribution type selected by the user. Appendix A includes a complete listing of the distribution functions and their parameters.

GETCDF: For the user-defined distribution type with the distribution parameters from DISPAR, this routines estimates the cumulative distribution function value, $F(x)$=P[Outcome $< x$] for given input x value.

FRACTL: For the user-defined distribution type with the distribution parameters from DISPAR, this routines estimates the fractile x corresponding to a specified input value of the probability p=$F(x)$=P[Outcome $< x$].

QDMOM: Uses Gaussian quadrature to estimate the first four moment of the theoretical fitted distribution. These can be compared with the sample moments from the data, as given by CALMOM, to verify the accuracy of the fitted model—and in the case of the higher moments not used in the original fitting, to test its accuracy.

The routines GETCDF and FRACTL, which supply general distribution functions and their inverses, may also be useful in other stand-alone applications; e.g., to create a distribution library for standard FORM/SORM or simulation analyses (Madsen et al, 1986), or for use with new Inverse FORM algorithms (Ude and Winterstein, 1996).

3 Input Format and Wave Height Example

3.1 Data Input

The file(s) containing data are read in free format, one datum per line. Non-numerical input are taken as comments and ignored. The first numerical value found is taken to be the duration of the database (in years). Remaining values (1 per line) are interpreted as data, and data are read until the end of the file is encountered.

We illustrate the input here through a simple example, involving significant wave height data. These are in fact 19 annual maximum H_S values, estimated by hindcast in a Southern North Sea location (Winterstein and Haver, 1991). These data are stored in appropriate format in the file `gumbel.dat`.

The contents of `gumbel.dat` are listed below. The first line reflects that 19 years are covered, and the remaining 19 lines contain the actual maximum H_S encountered in each of these 19 years. (In this input file the data are given in descending order; this is not required by the program.)

```
19.
9.66
9.44
9.18
9.17
8.85
8.79
8.60
8.58
8.54
8.49
8.09
8.08
8.06
7.47
```

```
7.42
7.41
7.31
7.16
6.92
```

Contents of gumbel.dat.

3.2 Runtime Input: Batch Mode

We seek to invoke FITS under the following conditions:

1. Results to be written to a file named gumbel1.out. (The distinction between lower- and upper-case letters in filenames is honored by Unix, and ignored in DOS.)
2. Distribution results are to be written for x (wave height) values ranging from XMIN.=5.0 to XMAX.=20.0 m, at increments of DX.=0.5 m.
3. There is only NFILES=1 database to be fit.
4. The wave height data are stored in a file named gumbel.dat.
5. The user desires to fit a Gumbel distribution (IDIST=5) to these data.

In this example FITS requires 5 lines of input, corresponding to the information given in the 5 items above. In this case the 5 input lines are as follows:

```
gumbel1.out      ;  Input line 1: File where output is to be written
5.0  20.0  0.5   ;  Input line 2: xmin, xmax, dx for writing dist results
1                ;  Input line 3: number of databases (input files)
gumbel.dat       ;  Input line 4: name of 1st database (input file)
5                ;  Input line 5: idist = dist type (idist=5 -> Gumbel)
```

(the ";" and following information is not read by FITS; these are given here to remind the user of the input definition). These 5 lines have been stored in a file named gumbel1.in. Thus, to execute FITS in a batch mode, the user can type (in either Unix or DOS environment) the following command:

```
fits < gumbel1.in
```

This will create (or overwrite) a corresponding file gumbel1.out, whose contents is discussed in the next section. While FITS is executing, the user will see prompts for terminal inputs; these can be simply ignored (or directed toward null device) in this batch mode operation.

3.3 Runtime Input: Interactive Mode

If the user simply types "FITS", he will be prompted for each input (the same 5 lines in this case) with interactive explanations. For example, before requesting the distribution index IDIST the program lists all available distribution types and associated IDIST values. Inputs with invalid formats are ignored. This interactive mode may be particularly useful for first-time users. (The text with input prompts is written to the logical unit IOERR, which is set to 0 in the driver program for FITS. The user can reset this if necessary.)

The following is a screen dump of the terminal input prompts and user's response. Lines beginning with ">" are input prompts generated by the program. Other lines are the user's response; in this case there are precisely 5 lines of response, as in the batch mode input file given previously.

```
>
> ** ENTER FILENAME WHERE OUTPUT WILL BE WRITTEN **
>
>       ENTER OUTPUT FILENAME:
gumbel1.out
>
> ** ENTER XMIN  = MIN X VALUE AT WHICH TO OUTPUT CDF
>            XMAX  = MAX X VALUE AT WHICH TO OUTPUT CDF
>              DX  = INTERVAL OF X VALS WHERE CDF IS OUTPUT
>     ALL THREE VALUES ON SAME LINE; E.G.
>
>        0.5 10.0 0.5
>
>     GIVES OUTPUT AT 20 X VALUES FROM 0.5 TO 10.0
>
>       ENTER XMIN,XMAX,DX:
5.0  20.0  0.5
>
> ** ENTER NFILES=NUMBER OF DATA SET FILES TO BE FIT
>     CURRENT OPTIONS: NFILES= 1  THROUGH   3
```

```
>
>       ENTER NFILES:
1
>
>         INPUT FILE NUMBER:   1
> ** ENTER FILENAME WHERE DATA ARE STORED,
>
>       ENTER INPUT FILENAME:
gumbel.dat
>
> ** ENTER IDIST =INDEX OF DISTRIBUTION TYPE TO BE FIT
>    CURRENT OPTIONS:
>                      IDIST =  1 ...  NORMAL
>                      IDIST =  2 ...  LOGNORMAL
>                      IDIST =  3 ...  EXPONENTIAL
>                      IDIST =  4 ...  WEIBULL
>                      IDIST =  5 ...  GUMBEL
>                      IDIST =  6 ...  SHIFTED EXPONENTIAL
>                      IDIST =  7 ...  SHIFTED WEIBULL
>
>       ENTER IDIST:
5
```

4 Output Format and Extreme Wave Height Example

The first output section of FITS provides summary statistics for each of the datafiles considered. These include (1) sample moments from the data, (2) predicted moments from the fitted distribution, and (3) underlying distribution parameters. Comparison of (1) and (2) can serve to verify the fit of low-order moments, and the agreement of higher moments (e.g., skewness and kurtosis) not used in the fitting can offer a rough goodness-of-fit test.

The second section of output gives distribution estimates for $G_i(x)$, the probability that a future outcome exceeds x as estimated from datafile i. It also reports the distribution of annual maxima, $G_{ann}(x)$, as estimated from Eq. 1. Note that all preceding output lines begin with "#", which is interpreted as a comment within the public-domain `gnuplot` plotting package. Thus the output file can be plotted directly with `gnuplot`.

The contents of the `gumbel1.out` output file follow here. The output confirms that, as intended, the fitted Gumbel model preserves the mean m_x=8.275m and standard deviation σ_x=0.816m estimated from the data. The Gumbel model tends to overestimate the higher moments, however; its predicted skewness 1.140 and kurtosis 5.40 varies notably from the values -0.053 and 1.91 estimated from the data.

This suggests that the Gumbel model may somewhat overestimate the chance of large wave heights in this case. This has also been observed previously for this data set (Winterstein and Haver, 1991), where a cubic distortion of the Gumbel model was introduced to capture this trend. Here we seek to model this effect by tail-fitting, within this population of annual maxima. Specifically, we also apply FITS to the same data with a shifted exponential model (IDIST=6). Two cases are considered, corresponding to lower-bounds of x_{low}=8.0m and 8.5m. The corresponding input and output files in these cases are `gumbel?.in` and `gumbel?.out`, with "?"=2 and 3. The input files show the additional input line required for shifted distributions, while the output (at least for x_{low}=8.5m) shows a somewhat reduced 100-year wave

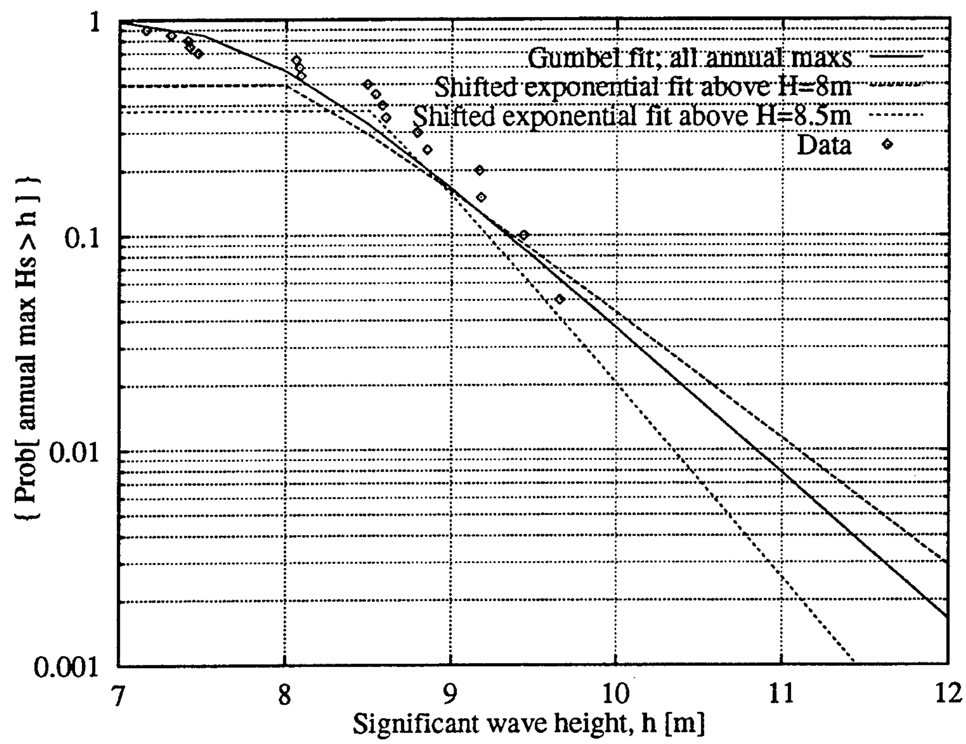

Figure 1: Estimating Annual Max H_S distribution.

height (10.4m vs 10.8m in the Gumbel case). These distribution fits are shown in Figure 1. They suggest that these shifted models can begin to capture trends in distribution tails, not available in moment-fits to the entire population.

A listing of `gumbel1.out` follows:

```
#
# RESULTS FOR FILE NUMBER:        1
#             FILE NAME   : gumbel.dat
#    DURATION OF DATABASE:     19.00
#      DIST TYPE SELECTED:   GUMBEL
#     NUMBER OF DATA USED:      19
#
#
#    MOMENTS FROM SAMPLE DATA     ( MEAN, SIGMA, SKEWNESS, KURTOSIS)
#   0.8275E+01  0.8186E+00 -0.5282E-01  0.1905E+01
#
#
#    MOMENTS FROM FITTED DIST     ( MEAN, SIGMA, SKEWNESS, KURTOSIS)
#   0.8275E+01  0.8186E+00  0.1140E+01  0.5400E+01
#
#
#    DISTRIBUTION PARAMETERS      (SEE DOCUMENTATION FOR DEFINITION)
#   0.8275E+01  0.8186E+00  0.1567E+01  0.7906E+01
#
#
# ** FITTED DISTRIBUTION RESULTS **
#
# P1 = Prob {Outcome > x} in database 1,
# P2 = Prob {Outcome > x} in database 2,
# ...
# Pc = Prob {Ann max > x} including all databases
#
#      X           P1    ....     Pc
  0.5000E+01  0.1000E+01  0.6321E+00
```

0.5500E+01	0.1000E+01	0.6321E+00
0.6000E+01	0.1000E+01	0.6321E+00
0.6500E+01	0.9999E+00	0.6321E+00
0.7000E+01	0.9840E+00	0.6262E+00
0.7500E+01	0.8489E+00	0.5721E+00
0.8000E+01	0.5783E+00	0.4392E+00
0.8500E+01	0.3260E+00	0.2782E+00
0.9000E+01	0.1649E+00	0.1520E+00
0.9500E+01	0.7905E-01	0.7600E-01
0.1000E+02	0.3692E-01	0.3625E-01
0.1050E+02	0.1704E-01	0.1690E-01
0.1100E+02	0.7822E-02	0.7791E-02
0.1150E+02	0.3581E-02	0.3575E-02
0.1200E+02	0.1638E-02	0.1636E-02
0.1250E+02	0.7486E-03	0.7483E-03
0.1300E+02	0.3421E-03	0.3420E-03
0.1350E+02	0.1563E-03	0.1563E-03
0.1400E+02	0.7141E-04	0.7141E-04
0.1450E+02	0.3263E-04	0.3263E-04
0.1500E+02	0.1491E-04	0.1491E-04
0.1550E+02	0.6810E-05	0.6810E-05
0.1600E+02	0.3111E-05	0.3111E-05
0.1650E+02	0.1421E-05	0.1421E-05
0.1700E+02	0.6494E-06	0.6494E-06
0.1750E+02	0.2967E-06	0.2967E-06
0.1800E+02	0.1356E-06	0.1356E-06
0.1850E+02	0.6193E-07	0.6193E-07
0.1900E+02	0.2830E-07	0.2829E-07
0.1950E+02	0.1293E-07	0.1293E-07
0.2000E+02	0.5906E-08	0.5906E-08

References

Madsen, H.O., S. Krenk, and N.C. Lind (1986). *Methods of structural safety,* Prentice-Hall, Inc., New Jersey.

Ude, T.C. and S.R. Winterstein (1996). *IFORM: An inverse-FORM routine to estimate response levels with specified return periods,* Rept. TN–3, Reliability of Marine Structures Program, Civil Eng. Dept., Stanford University.

Winterstein, S.R. and S. Haver (1991). Statistical uncertainties in wave heights and combined loads on offshore structures. *J. Offshore Mech. Arctic Eng.,* ASME, **114**(10), 1772–1790.

Winterstein, S.R., C.H. Lange, and S. Kumar (1994). *FITTING: A subroutine to fit four moment probability distributions to data,* Rept. RMS–14, Reliability of Marine Structures Program, Civil Eng. Dept., Stanford University.

A Specific Distribution Types and Functional Forms

Our purpose here is twofold. First, we seek to define the precise forms of probability distributions available in FITS. Second, we wish to indicate precisely how the distribution parameters are defined, and reported in FITS through the output of the subroutine DISPAR. This latter information may be useful if the user seeks to perform additional, off-line calculations based on the fitted distributions.

The following basic distribution types are currently available within FITS.

IDIST=1: **Normal Distribution.** The cumulative distribution function (CDF) is given by

$$P[\text{ Outcome } \leq x] = F(x) = \Phi(\frac{x - m_x}{\sigma_x}) \tag{2}$$

in terms of the standard normal CDF

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-u^2/2)du \tag{3}$$

In this case the subroutine DISPAR returns the parameters DPARM(1)=m_x and DPARM(2)=σ_x.

IDIST=2: **Lognormal Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = \Phi(\frac{\ln(x) - m_{\ln x}}{\sigma_{\ln x}}); \quad x \geq 0 \tag{4}$$

with $\Phi(x)$ again given by Eq. 3. In this case the subroutine DISPAR returns the parameters DPARM(1)=m_x, DPARM(2)=σ_x, DPARM(3)=$m_{\ln x}$, and DPARM(4)=$\sigma_{\ln x}$.

IDIST=3: **Exponential Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = 1 - \exp[-(\frac{x}{m_x})]; \quad x \geq 0 \tag{5}$$

In this case the subroutine DISPAR returns the parameters DPARM(1)=m_x and DPARM(2)=σ_x, and no additional parameters.

`IDIST`=4: **Weibull Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = 1 - \exp[-(\frac{x}{\beta})^{\alpha}]; \quad x \geq 0 \tag{6}$$

In this case the subroutine `DISPAR` returns the parameters DPARM(1)=m_x, DPARM(2)=σ_x, DPARM(3)=α, and DPARM(4)=β.

`IDIST`=5: **Gumbel Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = \exp\{-\exp[-\alpha(x-u)]\} \tag{7}$$

In this case the subroutine `DISPAR` returns the parameters DPARM(1)=m_x, DPARM(2)=σ_x, DPARM(3)=α, and DPARM(4)=u.

`IDIST`=6: **Shifted Exponential Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = 1 - \exp[-(\frac{x - x_{low}}{m_x})]; \quad x \geq x_{low} \tag{8}$$

As when `IDIST`=3, in this case the subroutine `DISPAR` returns the parameters DPARM(1)=m_x and DPARM(2)=σ_x, and no additional parameters.

`IDIST`=7: **Shifted Weibull Distribution.** The CDF is given by

$$P[\text{ Outcome } \leq x] = F(x) = 1 - \exp[-(\frac{x - x_{low}}{\beta})^{\alpha}]; \quad x \geq x_{low} \tag{9}$$

As when `IDIST`=4, in this case the subroutine `DISPAR` returns the parameters DPARM(1)=m_x, DPARM(2)=σ_x, DPARM(3)=α, and DPARM(4)=β.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*: 00-06-1996

2. REPORT DATE: June 1996

3. DATES COVERED *(From - To)*:

4. TITLE AND SUBTITLE: THE FITS ROUTINE: FITTING DISTRIBUTIONS TO MULTIPLE DATABASES AND ESTIMATING COMBINED EXTREMES

5a. CONTRACT NUMBER:

5b. GRANT NUMBER: N00014-96-1-0641

5c. PROGRAM ELEMENT NUMBER:

5d. PROJECT NUMBER:

5e. TASK NUMBER:

5f. WORK UNIT NUMBER:

6. AUTHOR(S): STEVEN R. WINTERSTEIN

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES):
RMS GROUP
S. R. WINTERSTEN, C.A. CORNELL
BLUME CENTER
STANFORD UNIVERSITY, CA 94305

8. PERFORMING ORGANIZATION REPORT NUMBER: RMS-19

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES):
OFFICE OF NAVAL RESEARCH
800 NORTH QUINCY ST.
ARLINGTON, VA 22217-4620
ATTN ROSHDY BARSOUM

10. SPONSOR/MONITOR'S ACRONYM(S):

11. SPONSORING/MONITORING AGENCY REPORT NUMBER:

12. DISTRIBUTION AVAILABILITY STATEMENT: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES:

14. ABSTRACT
THIS Report documents the usage of the routine FITS, which provides automated fits of various analytical, commonly used probability models from input data. The routine FITS is intended to provide robust fits (low moment) fits of simple, conventional distribution forms. For each data-base of interest, the routine estimates the distribution of annual maximum response, based on data, and duration over which they were recorded. It can also estimate maximal "events" ~~and~~ intended to reflect different statistical populations.

15. SUBJECT TERMS:

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPONE NUMBER (*Include area code*)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI-Std Z39-18